Week 1

What is a methodology anyway?

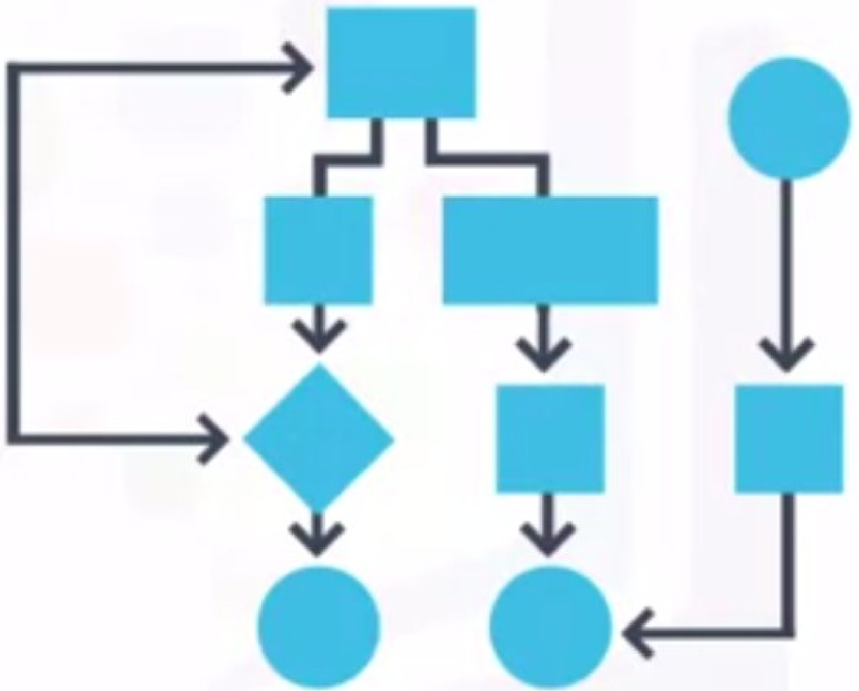
A methodology is a defined way of....

meth·od·ol·o·gy

noun

noun: **methodology**; plural noun: **methodologies**

1 a system of methods used in a particular area of study or activity.
"a methodology for investigating the concept of focal points"



It's important to consider it because all too often there is a temptation to bypass

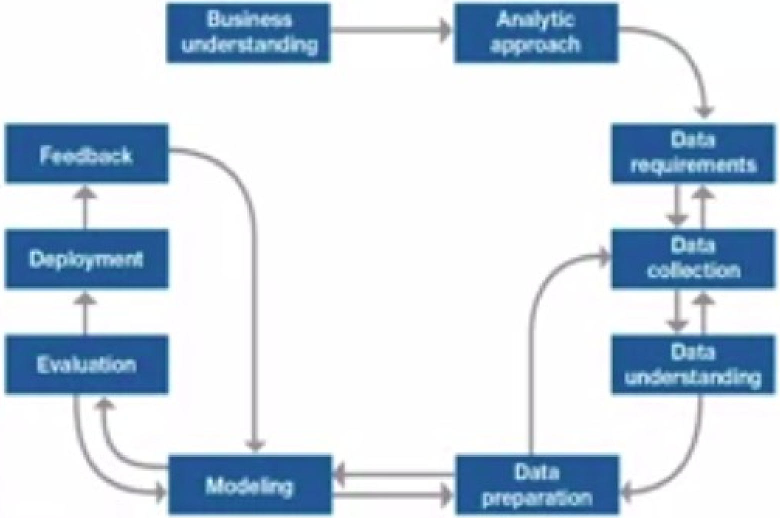
Methodology by John Rollins based on CRISP-DM

John Rollins
Data Scientist, IBM Analytics, IBM

John B. Rollins, Ph.D., P.E., is a Data Scientist, IBM Analytics, IBM. Prior to joining IBM Netezza, he was an engineering consultant, professor and researcher. He has authored many patents, papers and books. He holds doctoral degrees in economics and petroleum engineering and is a registered professional engineer in Texas.

a seasoned and senior data scientist currently practising at IBM. This course is built on

In a nutshell...



The **Data Science Methodology** aims to answer the following 10 questions in this prescribed sequence:

From problem to approach:

1. *What is the problem that you are trying to solve?*
2. *How can you use data to answer the question?*

Working with the data:

3. *What data do you need to answer the question?*
4. *Where is the data coming from (identify all sources) and how will you get it?*
5. *Is the data that you collected representative of the problem to be solved?*
6. *What additional work is required to manipulate and work with the data?*

Deriving the answer:

7. *In what way can the data be visualized to get to the answer that is required?*
8. *Does the model used really answer the initial question or does it need to be adjusted?*
9. *Can you put the model into practice?*
10. *Can you get constructive feedback into answering the question?*

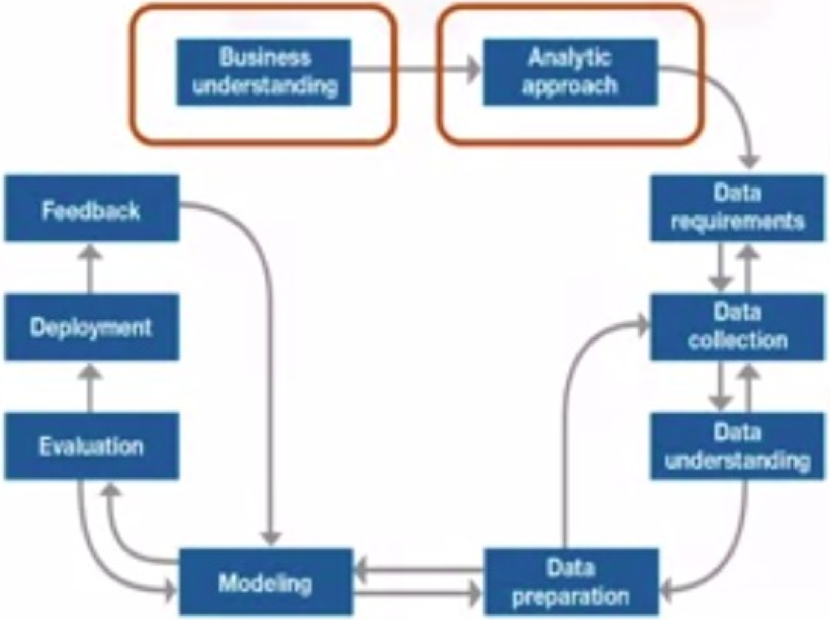
Course structure

- **Module 1: From Problem to Approach**
  - Business Understanding – Concepts & Case Study
  - Analytic Approach – Concepts & Case Study
  - Hands-on Lab & Review
- **Module 2: From Requirements to Collection**
  - Data Requirements – Concepts & Case Study
  - Data Collection – Concepts & Case Study
  - Hands-on Lab & Review
- **Module 3: From Understanding to Preparation**
  - Data Understanding – Concepts & Case Study
  - Data Preparation – Concepts
  - Data Preparation – Case Study
  - Hands-on Lab & Review
- **Module 4: From Modeling to Evaluation**
  - Modeling – Concepts
  - Modeling – Case Study
  - Evaluation – Concepts & Case Study
  - Hands-on Lab & Review
- **Module 5: From Deployment to Feedback**
  - Deployment – Concepts & Case Study
  - Feedback – Concepts & Case Study
  - Hands-on Lab & Review

Glossary of Data Science Terms

- analytic approach
- analytics
- cohort
- cohort study
- comorbidities
- congestive heart failure (CHF)
- CRISP-DM
- data analysis
- data cleansing
- data science
- data scientist
- decision tree
- decision tree classification
- descriptive modeling
- descriptive statistics
- domain knowledge
- dominating decision rule
- histogram
- hospital readmission
- Iterative process > Iteration
- machine learning
- mean
- median
- methodology
- model > conceptual model
- pairwise comparison (correlation)
- patient cohort
- pattern
- predictive modeling
- predictors
- ROC curve
- standard deviation
- statistics
- structured data > data model
- text analysis > data mining
- training set
- univariate
- unstructured data
- visualization techniques

From Understanding to Approach



Business understanding

- *What is the problem that you are trying to solve?*

Analytic approach

- *How can you use data to answer the question?*

Case Study – What are the goals & objectives?

Define the GOALS

- To provide quality care without increasing costs

Define the OBJECTIVES

- To review the process to identify inefficiencies

Pick analytic approach based on type of question

Descriptive
- Current status

Diagnostic (Statistical Analysis)
- What happened?
- Why is this happening?

Predictive (Forecasting)
- What if these trends continue?
- What will happen next?

Prescriptive
- How do we solve it?

What are the types of questions?

If the question is to determine probabilities of an action

- Use a Predictive model

If the question is to show relationships

- Use a descriptive model

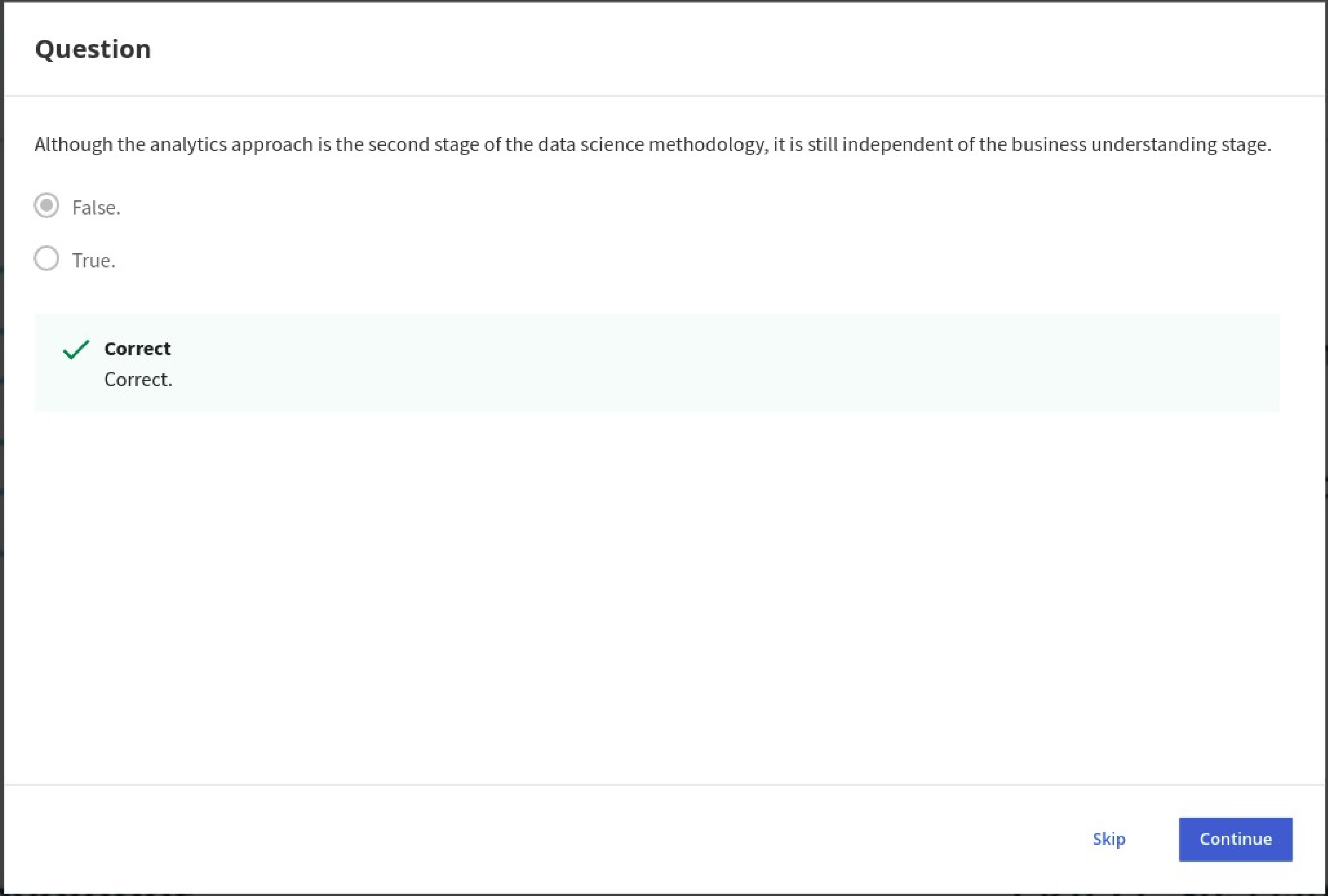
If the question requires a yes/no answer

- Use a classification model

Analytic approach

- *How can you use data to answer the question?*

- The correct approach depends on business requirements for the model



Question

Although the analytics approach is the second stage of the data science methodology, it is still independent of the business understanding stage.

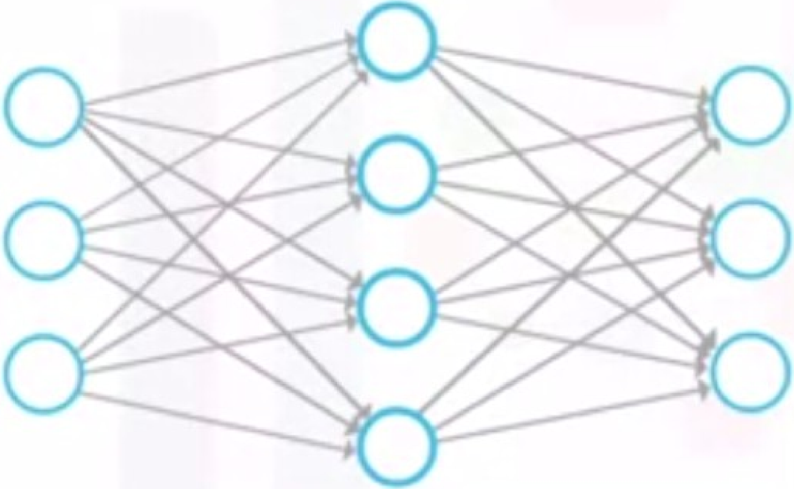
● False.

○ True.

✓ **Correct**

Correct.

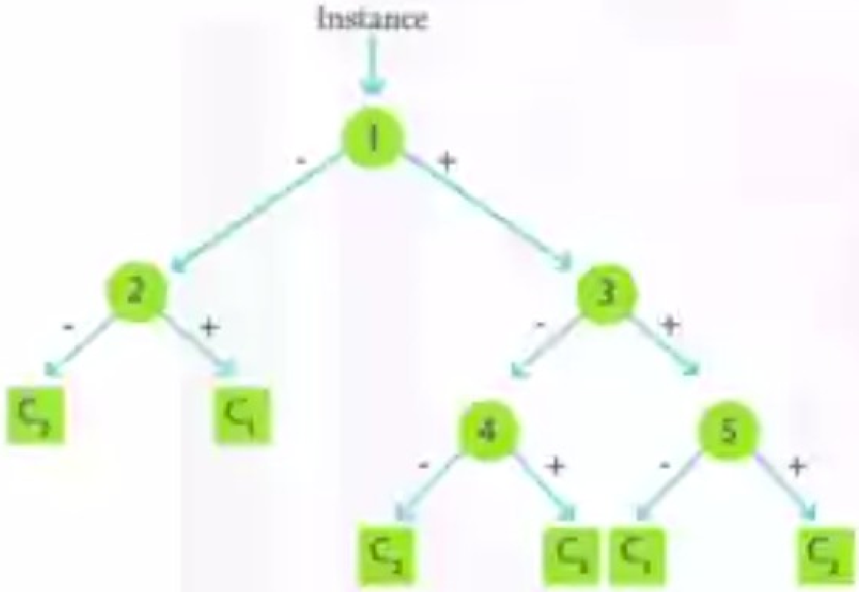
Will machine learning be utilized?



Machine Learning

- Learning without being explicitly programmed
- Identifies relationships and trends in data that might otherwise not be accessible or identified
- Uses clustering association approaches

Case Study – Decision tree classification selected!



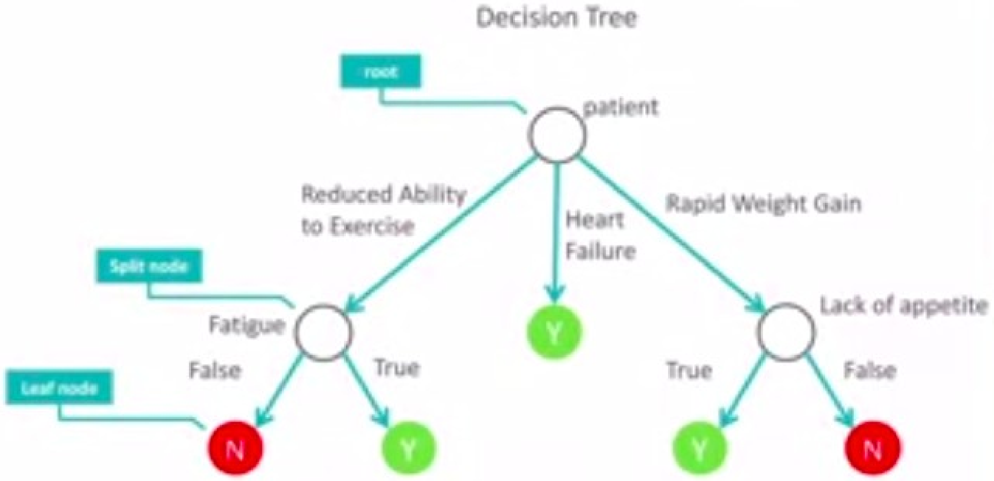
Predictive model

- To predict an outcome

Decision tree classification

- Categorical outcome
- Explicit "decision path" showing conditions leading to high risk
- Likelihood of classified outcome
- Easy to understand and apply

Case Study – Example of decision tree classification

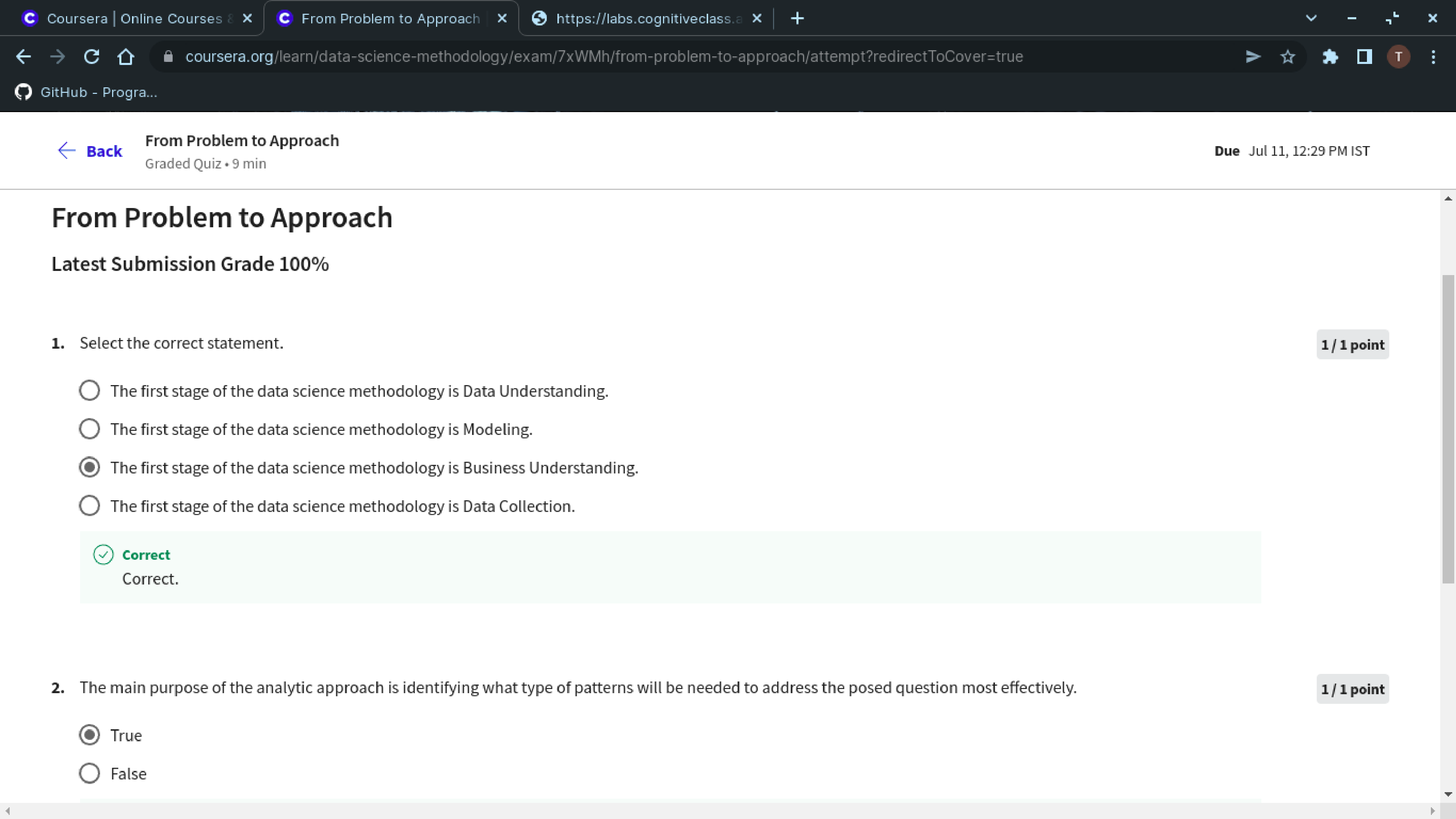


Predictive model

- To predict an outcome

Decision tree classification

- Categorical outcome
- Explicit "decision path" showing conditions leading to high risk
- Likelihood of classified outcome
- Easy to understand and apply


Coursera | Online Courses &
From Problem to Approach
https://labs.cognitiveclass.a
coursera.org/learn/data-science-methodology/exam/7xWMh/from-problem-to-approach/attempt?redirectToCover=true
GitHub - Progra...
Back

From Problem to Approach
Graded Quiz • 9 min

Due Jul 11, 12:29 PM IST

From Problem to Approach

Latest Submission Grade 100%

1. Select the correct statement. **1 / 1 point**

○ The first stage of the data science methodology is Data Understanding.

○ The first stage of the data science methodology is Modeling.

● The first stage of the data science methodology is Business Understanding.

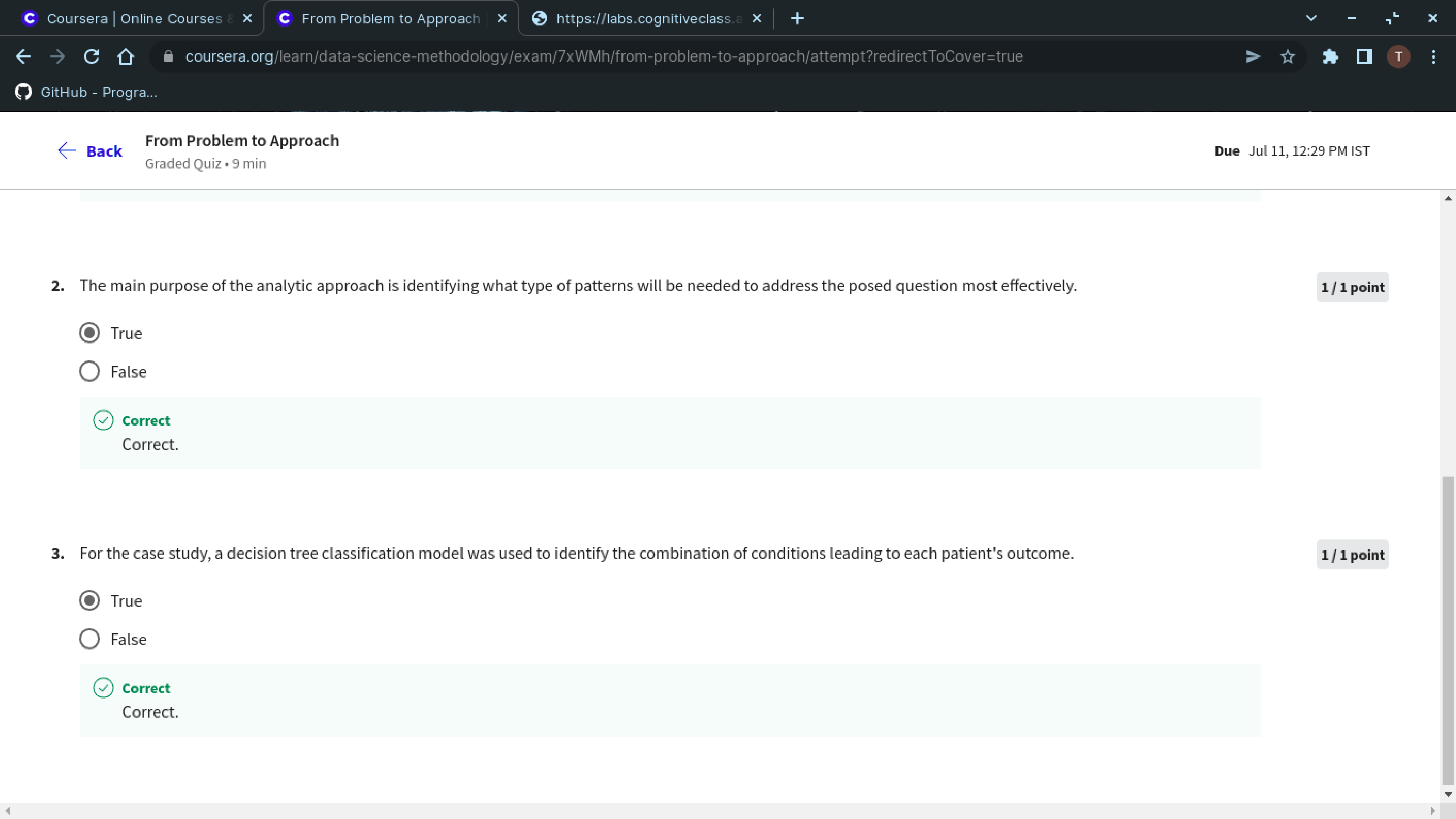
○ The first stage of the data science methodology is Data Collection.

✓ **Correct**
Correct.

2. The main purpose of the analytic approach is identifying what type of patterns will be needed to address the posed question most effectively. **1 / 1 point**

● True

○ False


Coursera | Online Courses &
From Problem to Approach
https://labs.cognitiveclass.a
coursera.org/learn/data-science-methodology/exam/7xWMh/from-problem-to-approach/attempt?redirectToCover=true
GitHub - Progra...
Back

From Problem to Approach

Graded Quiz • 9 min

Due Jul 11, 12:29 PM IST

2. The main purpose of the analytic approach is identifying what type of patterns will be needed to address the posed question most effectively.

1 / 1 point

● True

○ False

Correct

Correct.

3. For the case study, a decision tree classification model was used to identify the combination of conditions leading to each patient's outcome.

1 / 1 point

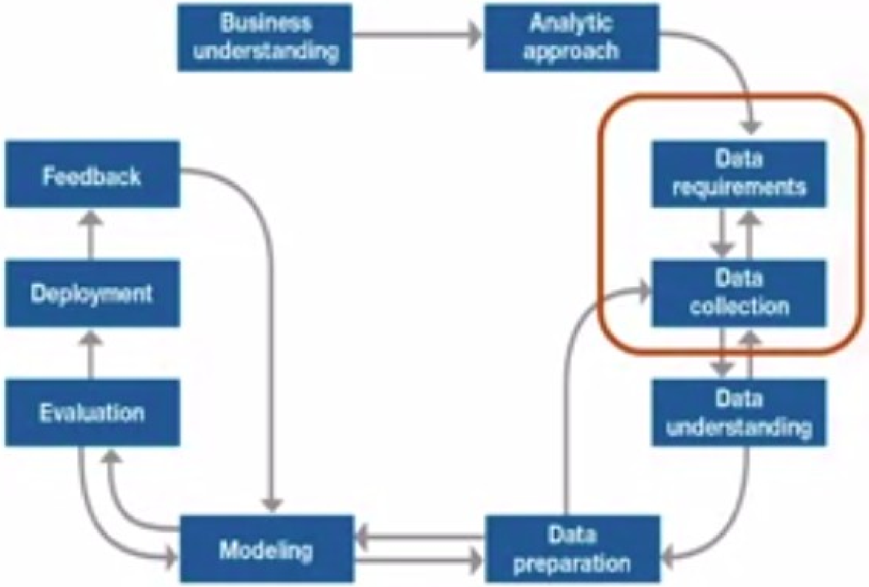
● True

○ False

Correct

Correct.

From Requirements to Collection



Data Requirements

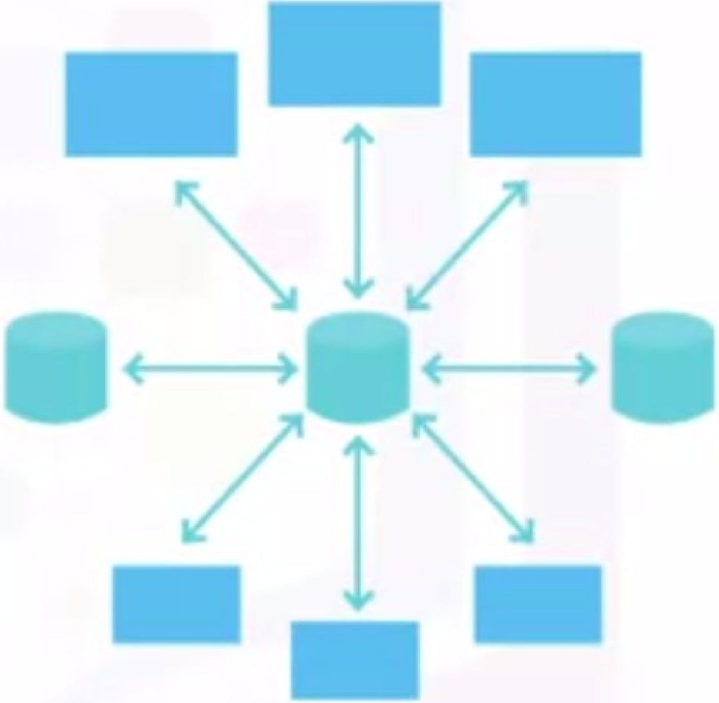
- *What are data requirements?*

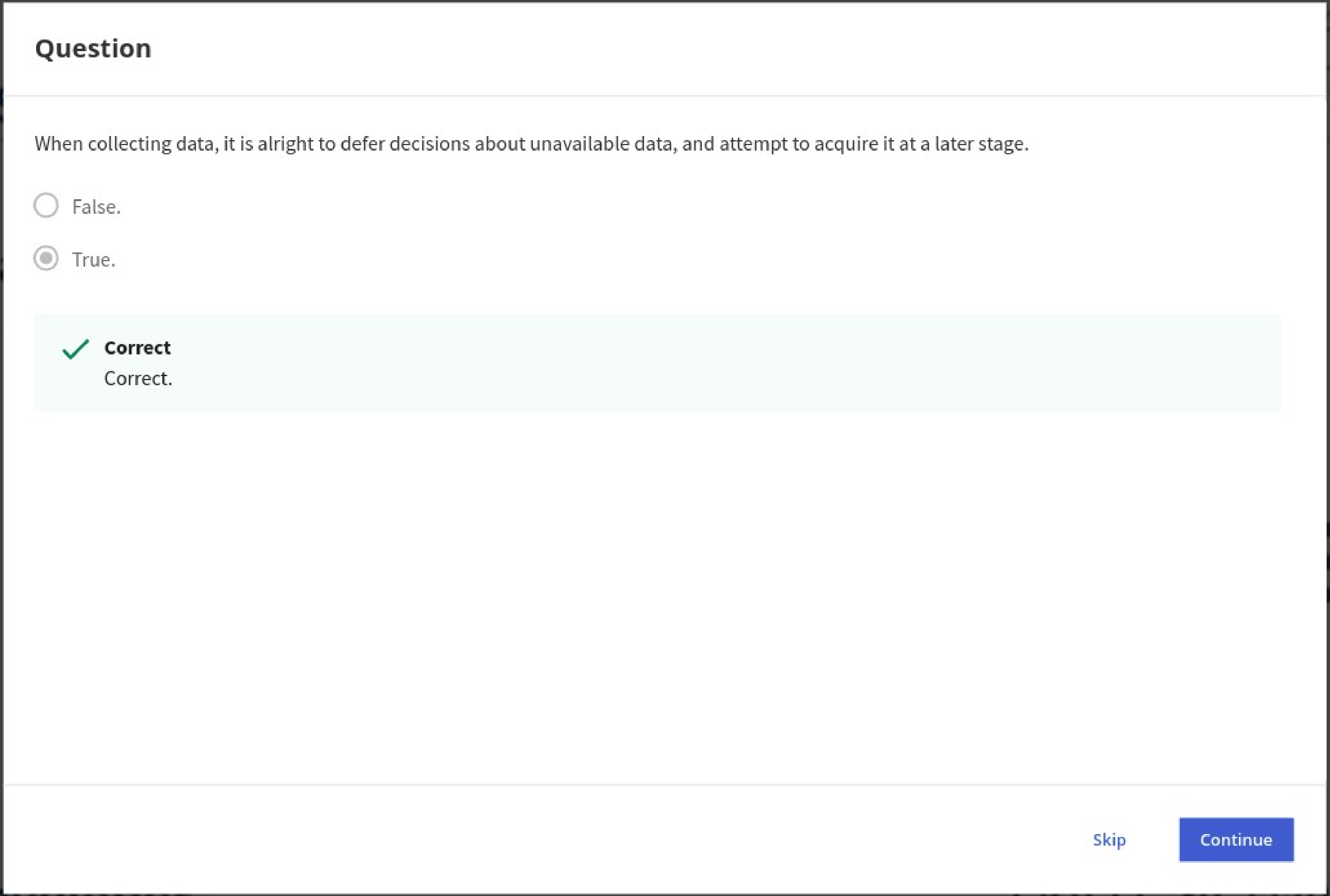
Data Collection

- *What occurs during data collection?*

Case Study – Gathering available data

- Available data sources
  - Corporate data warehouse (single source of medical & claims, eligibility, provider and member information)
  - In-patient record system
  - Claim payment system
  - Disease management program information





Question

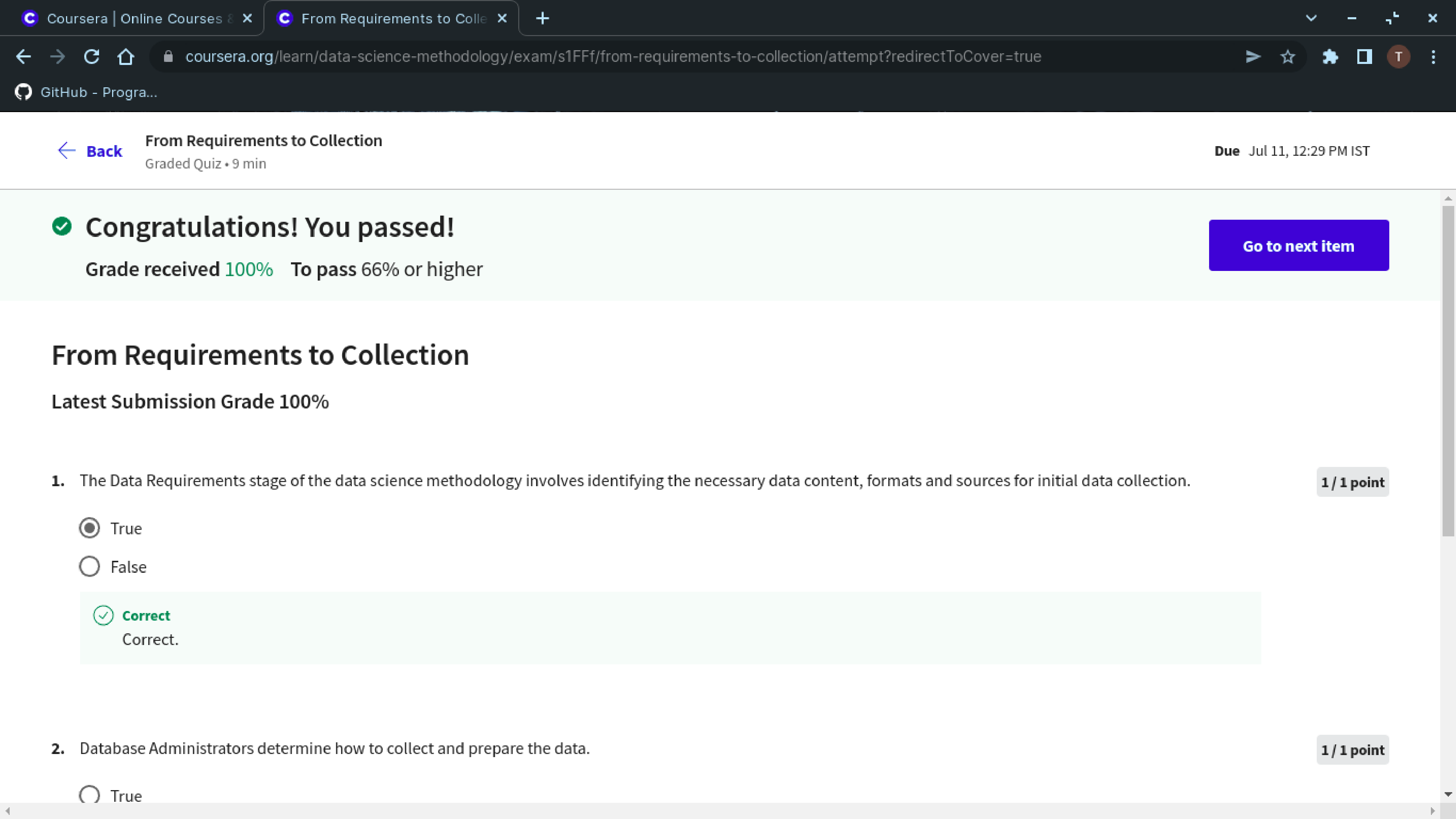
When collecting data, it is alright to defer decisions about unavailable data, and attempt to acquire it at a later stage.

○ False.

● True.

✓ **Correct**

Correct.


Coursera | Online Courses &
From Requirements to Colle
coursera.org/learn/data-science-methodology/exam/s1FFf/from-requirements-to-collection/attempt?redirectToCover=true
GitHub - Progra...
Back
Go to next item

From Requirements to Collection

Graded Quiz • 9 min

Due Jul 11, 12:29 PM IST

Congratulations! You passed!

Grade received 100% **To pass** 66% or higher

From Requirements to Collection

Latest Submission Grade 100%

1. The Data Requirements stage of the data science methodology involves identifying the necessary data content, formats and sources for initial data collection. **1 / 1 point**

● True

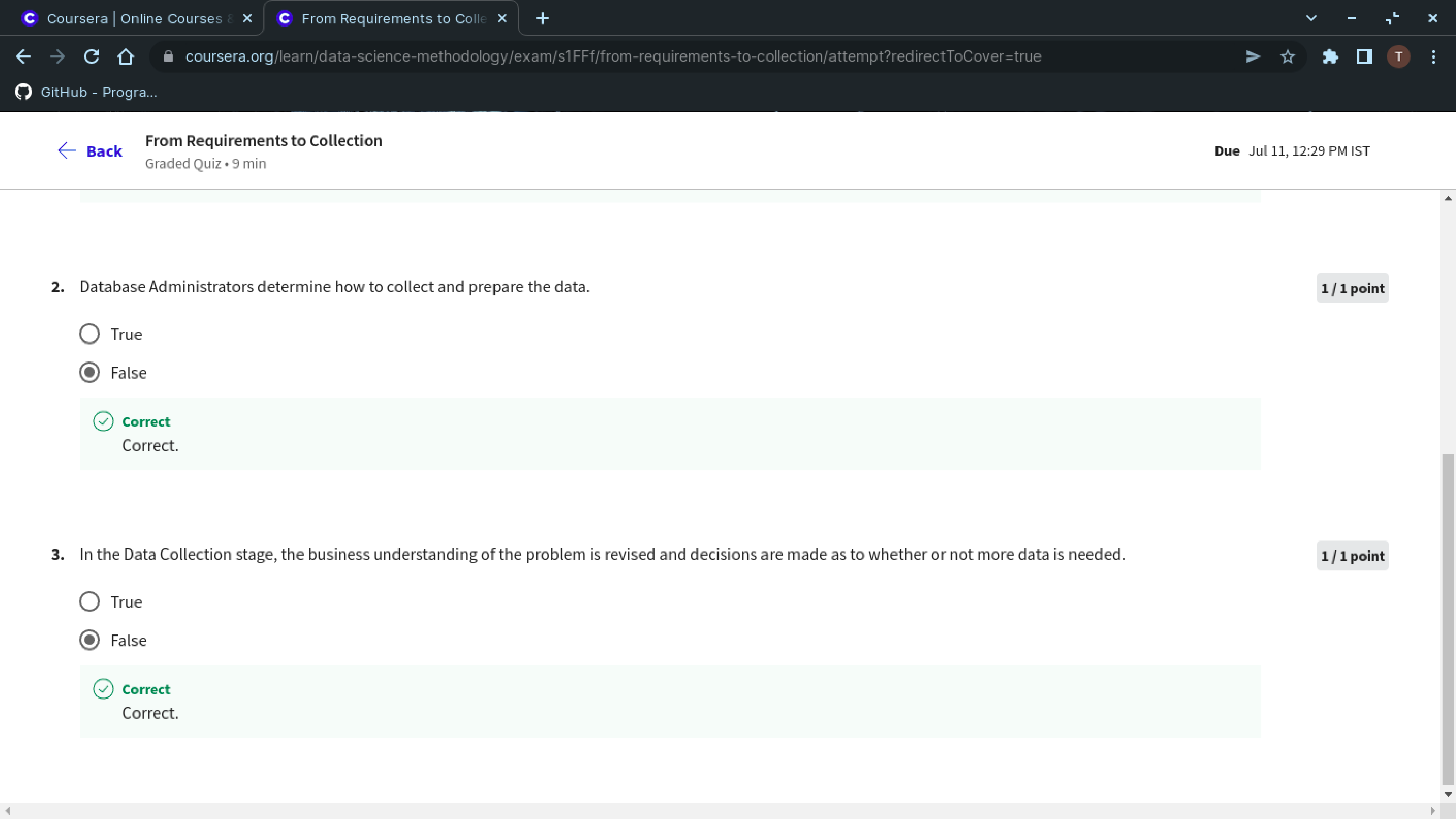
○ False

Correct

Correct.

2. Database Administrators determine how to collect and prepare the data. **1 / 1 point**

○ True


Coursera | Online Courses &
From Requirements to Colle
coursera.org/learn/data-science-methodology/exam/s1FFf/from-requirements-to-collection/attempt?redirectToCover=true
GitHub - Progra...
Back

From Requirements to Collection

Graded Quiz • 9 min

Due Jul 11, 12:29 PM IST

2. Database Administrators determine how to collect and prepare the data. **1 / 1 point**

○ True

● False

Correct

Correct.

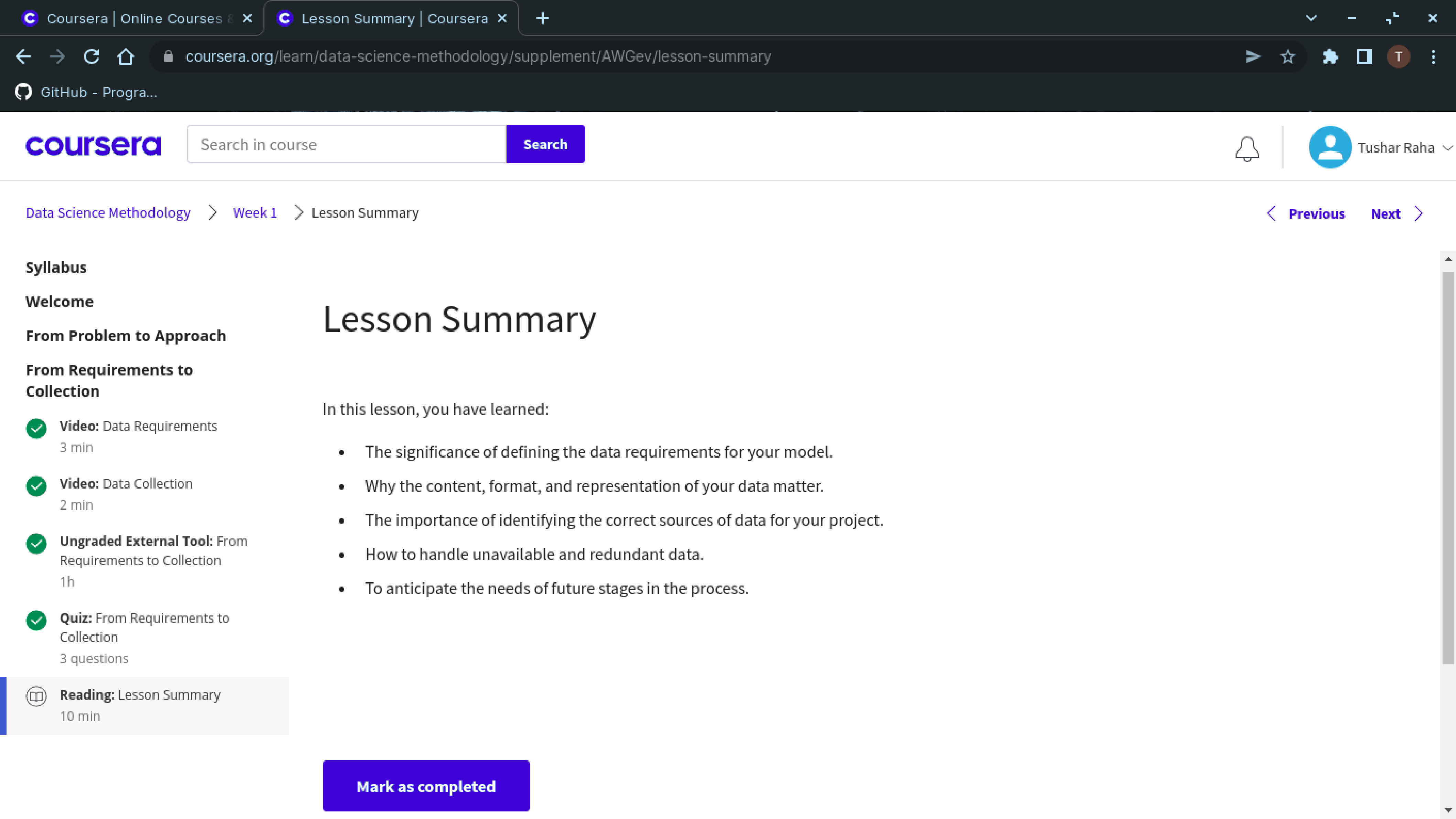
3. In the Data Collection stage, the business understanding of the problem is revised and decisions are made as to whether or not more data is needed. **1 / 1 point**

○ True

● False

Correct

Correct.



Coursera | Online Courses &
Lesson Summary | Coursera
coursera.org/learn/data-science-methodology/supplement/AWGev/lesson-summary
GitHub - Progra...
coursera
Search in course
Search
Tushar Raha
Previous
Next

Data Science Methodology > Week 1 > Lesson Summary

Syllabus

Welcome

From Problem to Approach

From Requirements to Collection

- **Video:** Data Requirements
  3 min
- **Video:** Data Collection
  2 min
- **Ungraded External Tool:** From Requirements to Collection
  1h
- **Quiz:** From Requirements to Collection
  3 questions
- **Reading:** Lesson Summary
  10 min

Lesson Summary

In this lesson, you have learned:

- The significance of defining the data requirements for your model.
- Why the content, format, and representation of your data matter.
- The importance of identifying the correct sources of data for your project.
- How to handle unavailable and redundant data.
- To anticipate the needs of future stages in the process.

Mark as completed